Introduction of VCDE Interoperability Review for Protein Information Resource (PIR)

caBIG™ VCDE Workspace Monthly Teleconference
November 3, 2005

Baris Ethem Suzek
Georgetown University - Lombardi Cancer Center – PIR
bes23@georgetown.edu

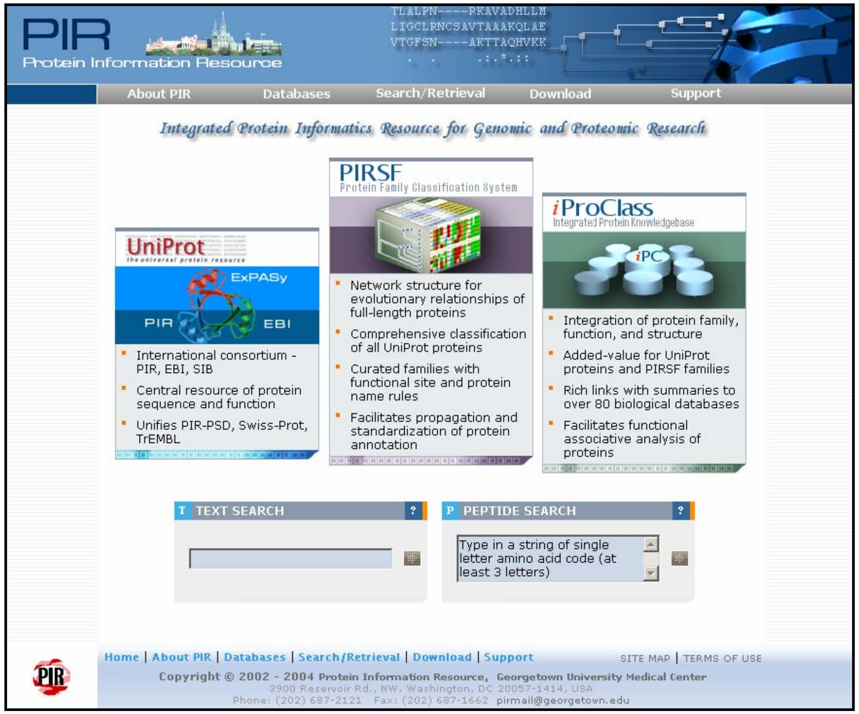
PIR
Protein Information Resource
A UniProt CONSORTIUM MEMBER

Outline

- Introduction
- Overview of Grid-Enablement of PIR
- Data Model
- Semantic Annotation
- Using PIR Grid Service
- Discussion

Introduction

- Protein Information Resource (PIR): Integrated Protein Informatics Resource for Genomic/Proteomic Research



Integrated Protein Informatics Resource for Genomic and Proteomic Research

UniProt

- International consortium - PIR, EBI, SIB
- Central resource of protein sequence and function
- Unifies PIR-PSD, Swiss-Prot, TrEMBL

PIRSF

- Network structure for evolutionary relationships of full-length proteins
- Comprehensive classification of all UniProt proteins
- Curated families with functional site and protein name rules
- Facilitates propagation and standardization of protein annotation

iProClass

- Integration of protein family, function, and structure
- Added-value for UniProt proteins and PIRSF families
- Rich links with summaries to over 80 biological databases
- Facilitates functional associative analysis of proteins

Type in a string of single letter amino acid code (at least 3 letters)

Home | About PIR | Databases | Search/Retrieval | Download | Support   SITE MAP | TERMS OF USE

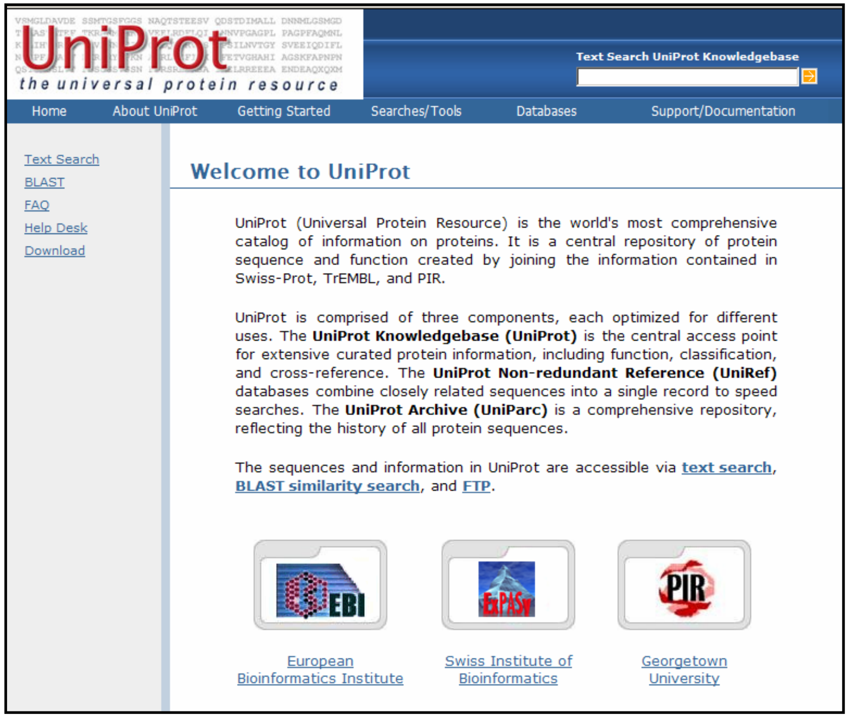
Copyright © 2002 - 2004 Protein Information Resource, Georgetown University Medical Center
3900 Reservoir Rd., NW, Washington, DC 20057-1414, USA
Phone: (202) 687-2121 Fax: (202) 687-1662 pirmail@georgetown.edu

- UniProt Universal Protein Resource: Central Resource of Protein Sequence and Function
- PIRSF Family Classification System: Protein Classification and Functional Annotation
- iProClass Integrated Protein Knowledgebase: Data Integration and Functional Analysis

http://pir.georgetown.edu

Introduction

- UniProt: Universal Protein Resource - Central Resource of Protein Sequence and Function



Welcome to UniProt

UniProt (Universal Protein Resource) is the world's most comprehensive catalog of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR.

UniProt is comprised of three components, each optimized for different uses. The **UniProt Knowledgebase (UniProt)** is the central access point for extensive curated protein information, including function, classification, and cross-reference. The **UniProt Non-redundant Reference (UniRef)** databases combine closely related sequences into a single record to speed searches. The **UniProt Archive (UniParc)** is a comprehensive repository, reflecting the history of all protein sequences.

The sequences and information in UniProt are accessible via text search, BLAST similarity search, and FTP.

European Bioinformatics Institute

Swiss Institute of Bioinformatics

Georgetown University

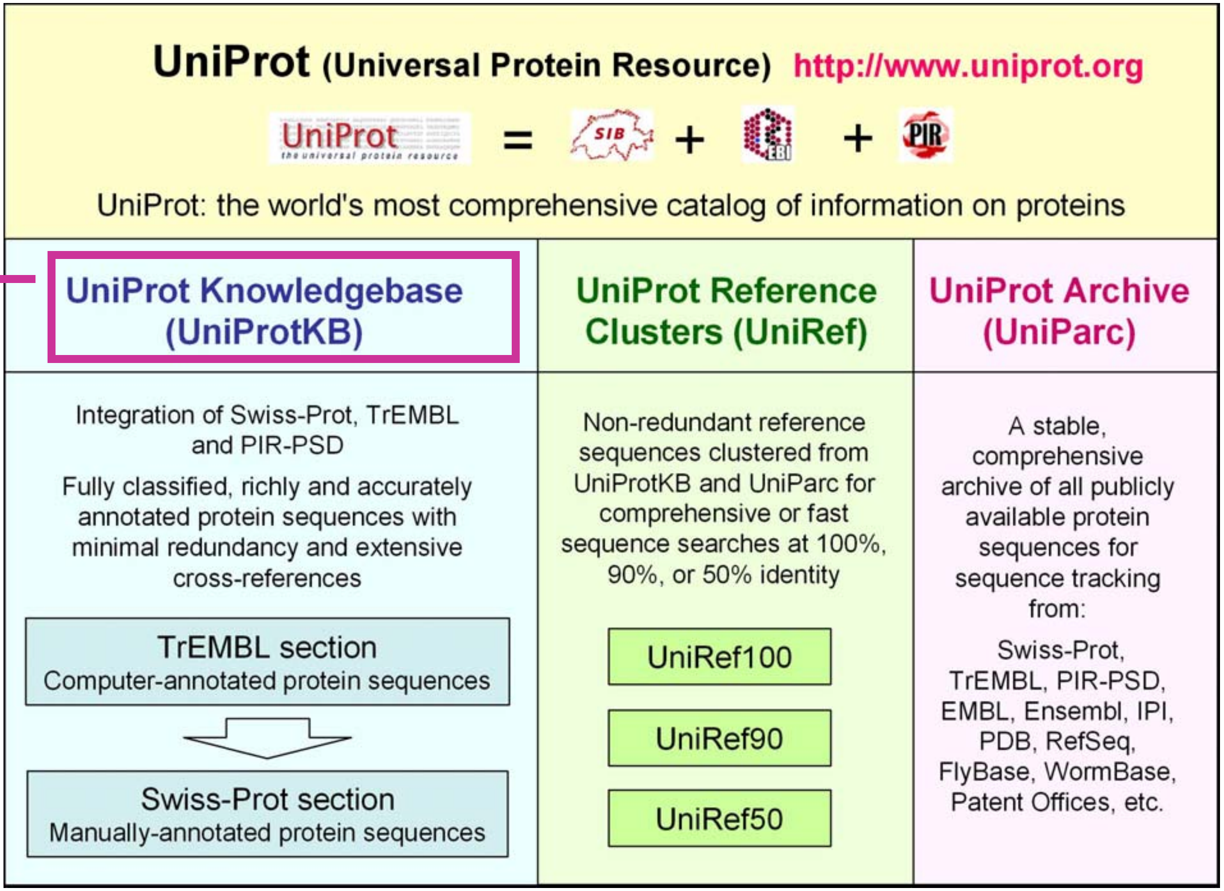
- International Consortium
  - PIR at GUMC
  - European Bioinformatics Institute (EBI)
  - Swiss Institute of Bioinformatics (SIB)
- Unifies PIR-PSD, Swiss-Prot, TrEMBL Protein Sequence Databases

http://www.uniprot.org

Introduction

- UniProt Databases

Primary data source for Grid-Enablement of PIR



UniProt (Universal Protein Resource) http://www.uniprot.org

UniProt = SIB + EBI + PIR

UniProt: the world's most comprehensive catalog of information on proteins

UniProt Knowledgebase (UniProtKB)

Integration of Swiss-Prot, TrEMBL and PIR-PSD

Fully classified, richly and accurately annotated protein sequences with minimal redundancy and extensive cross-references

TrEMBL section
Computer-annotated protein sequences

Swiss-Prot section
Manually-annotated protein sequences

UniProt Reference Clusters (UniRef)

Non-redundant reference sequences clustered from UniProtKB and UniParc for comprehensive or fast sequence searches at 100%, 90%, or 50% identity

UniRef100

UniRef90

UniRef50

UniProt Archive (UniParc)

A stable, comprehensive archive of all publicly available protein sequences for sequence tracking from:

Swiss-Prot, TrEMBL, PIR-PSD, EMBL, Ensembl, IPI, PDB, RefSeq, FlyBase, WormBase, Patent Offices, etc.

Project Overview

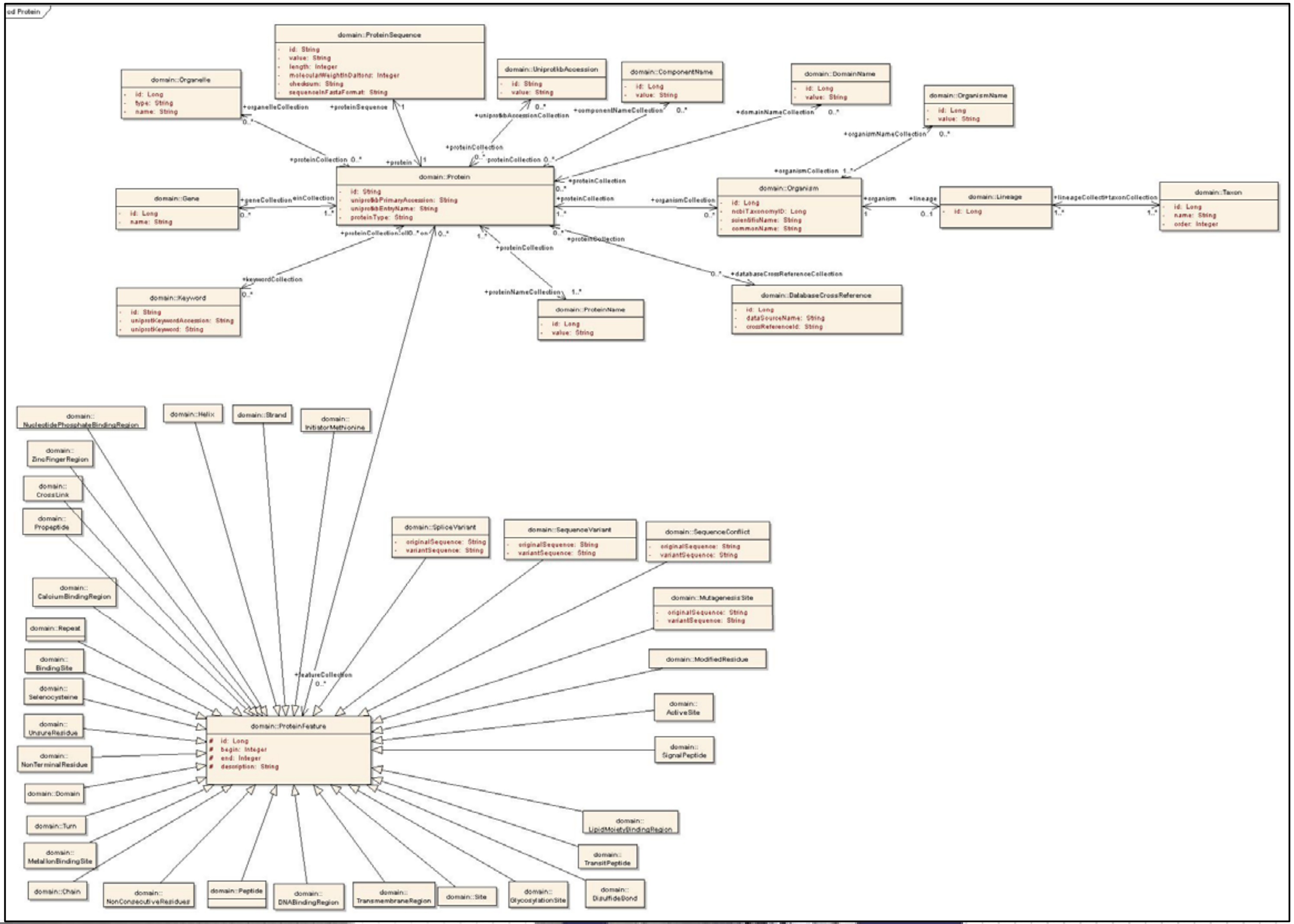
- Goal: Providing methods to query and retrieve protein related information for the cancer research community
- Grid-Enablement of PIR project is a data service
- All the objects in our model exposed to caGRID as of August 1st
- API is generated using caCORE SDK 1.0.3 like caBIO
- Example queries:
  - Find the proteins for the gene “BRCA2” (Breast Cancer Gene 2)
  - Find all the proteins that contain the domain BRCA2 repeat (PFAM:PF00634, a domain in Breast cancer type 2 susceptibility protein)
  - ID mapping: Find all the database cross-references from various databases corresponding to RefSeq Accession NP_009225

Data Model

- In collaboration with Lewis Frey and George Komatsoulis
- Considerations:
  - Scientific meaning– Don’t do record modeling
  - Use cases – Consider search criteria objects
  - caCORE SDK constraints – Consider naming conventions, “id” attribute constraints, supported collection types. e.g. “List” is not supported
  - Data related constraints – Include only associations or objects based on your data. e.g. Gene to Protein, but not Protein to DNASequence
  - Semantics – Express semantics and avoid using type attributes. e.g. ProteinFeature subclasses, Lineage
  - Other projects – Review caCORE/caBIO models

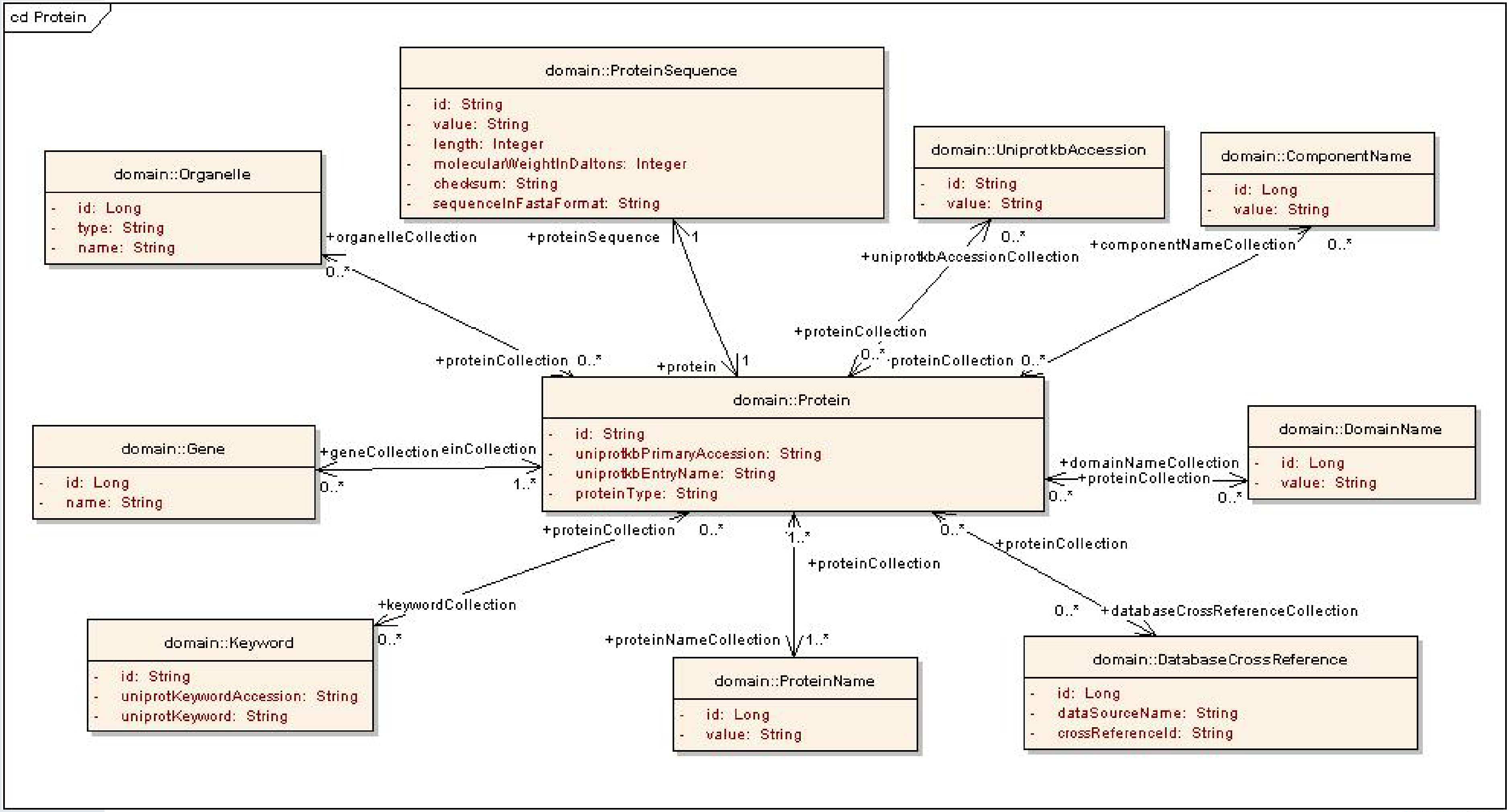
Data Model

- Total 48 objects , 51 attributes



Data Model

- Protein/Gene related objects



cd Protein

domain::ProteinSequence
- id: String
- value: String
- length: Integer
- molecularWeightInDaltons: Integer
- checksum: String
- sequenceInFastaFormat: String

domain::Organelle
- id: Long
- type: String
- name: String

domain::UniprotkbAccession
- id: String
- value: String

domain::ComponentName
- id: Long
- value: String

domain::Protein
- id: String
- uniprotkbPrimaryAccession: String
- uniprotkbEntryName: String
- proteinType: String

domain::Gene
- id: Long
- name: String

domain::DomainName
- id: Long
- value: String

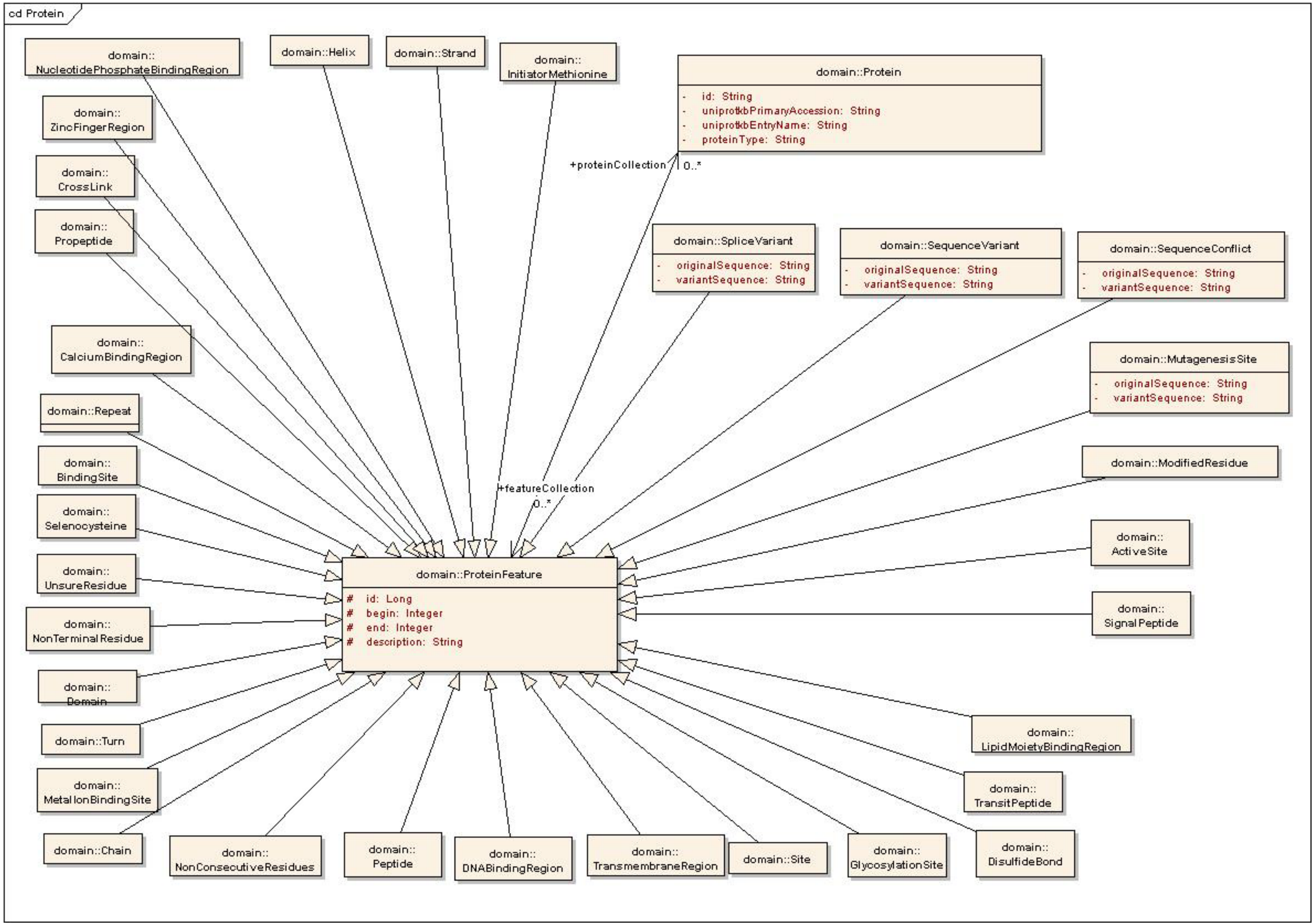
domain::Keyword
- id: String
- uniprotKeywordAccession: String
- uniprotKeyword: String

domain::ProteinName
- id: Long
- value: String

domain::DatabaseCrossReference
- id: Long
- dataSourceName: String
- crossReferenceId: String

Data Model

- Annotation related objects: Protein Features



cd Protein

domain::Protein
- id: String
- uniprotkbPrimaryAccession: String
- uniprotkbEntryName: String
- proteinType: String

+proteinCollection 0..*

+featureCollection 0..*

domain::ProteinFeature
# id: Long
# begin: Integer
# end: Integer
# description: String

domain::SpliceVariant
- originalSequence: String
- variantSequence: String

domain::SequenceVariant
- originalSequence: String
- variantSequence: String

domain::SequenceConflict
- originalSequence: String
- variantSequence: String

domain::MutagenesisSite
- originalSequence: String
- variantSequence: String

domain::NucleotidePhosphateBindingRegion

domain::Helix

domain::Strand

domain::InitiatorMethionine

domain::ZincFingerRegion

domain::CrossLink

domain::Propeptide

domain::CalciumBindingRegion

domain::Repeat

domain::BindingSite

domain::Selenocysteine

domain::UnsureResidue

domain::NonTerminalResidue

domain::Domain

domain::Turn

domain::MetalIonBindingSite

domain::Chain

domain::NonConsecutiveResidues

domain::Peptide

domain::DNABindingRegion

domain::TransmembraneRegion

domain::Site

domain::GlycosylationSite

domain::DisulfideBond

domain::TransitPeptide

domain::LipidMoietyBindingRegion

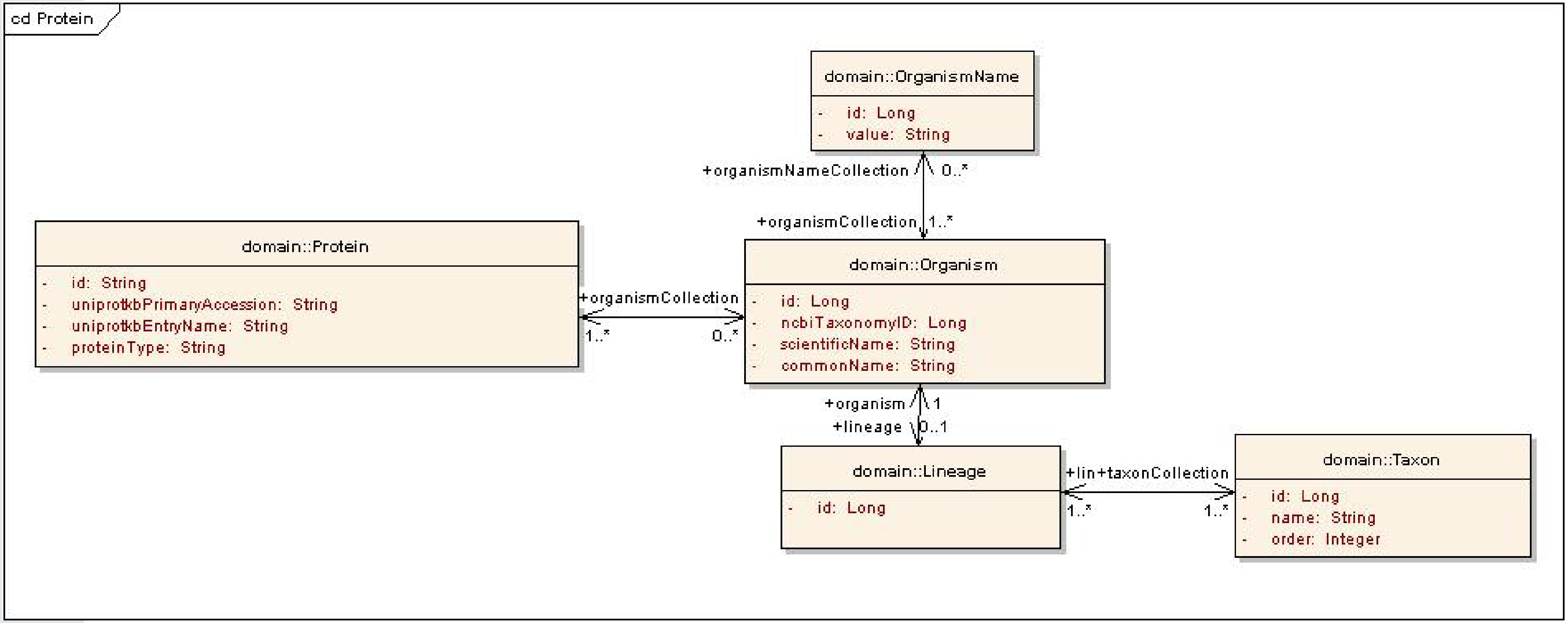
domain::SignalPeptide

domain::ActiveSite

domain::ModifiedResidue

Data Model

- Taxonomy related objects (Proposed as Taxonomy CDE)



cd Protein

domain::OrganismName
- id: Long
- value: String

+organismNameCollection 0..*

+organismCollection 1..*

domain::Protein
- id: String
- uniprotkbPrimaryAccession: String
- uniprotkbEntryName: String
- proteinType: String

+organismCollection
1..*
0..*

domain::Organism
- id: Long
- ncbiTaxonomyID: Long
- scientificName: String
- commonName: String

+organism 1
+lineage 0..1

domain::Lineage
- id: Long

+lin+taxonCollection
1..*
1..*

domain::Taxon
- id: Long
- name: String
- order: Integer

Semantic Annotation

- 149 concepts are used
- Loaded to caDSR production server on August 8
- Example: Gene.name
  - Property:
    - C42614: Name: The words or language units by which a thing is known.
  - PropertyQualifier1:
    - C43568: Gene_Symbol: A unique gene name approved by an organism specific nomenclature committee.

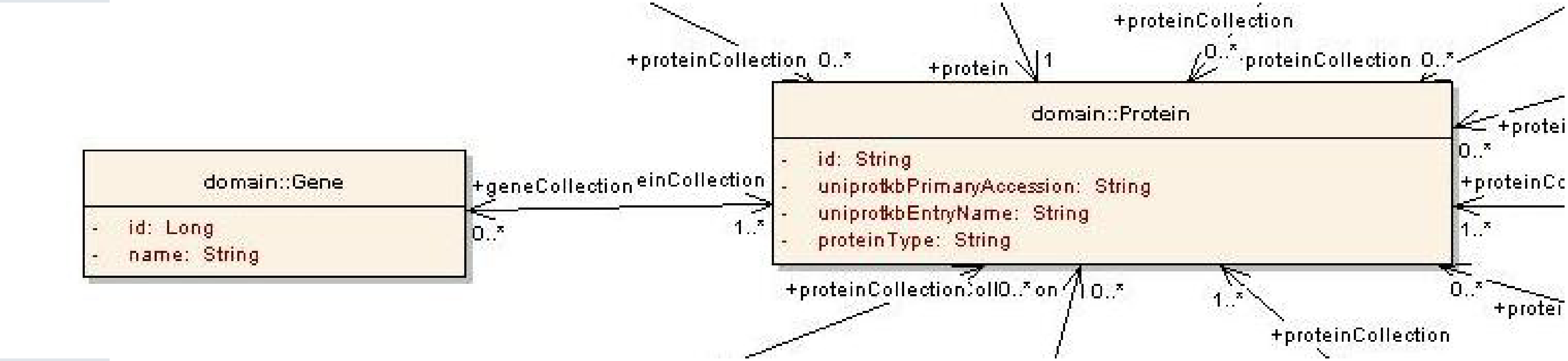
Semantic Annotation

- Example: Protein.uniprotkbPrimaryAccession
  - Property:
    - C15402: Accession_Number: A control number unique to an object, used to identify it among the other objects in a collection.
  - PropertyQualifier1:
    - C4785: UniProt_KB: The UniProt Knowledgebase (UniProtKB), a product of the UniProt consortium, provides a central database of protein sequences with accurate, consistent, rich sequence and functional annotation. The UniProt Knowledgebase consists of two sections: Swiss-Prot - a section containing manually-annotated records with information extracted from literature and curator-evaluated computational analysis, and TrEMBL - a section with computationally analyzed records that await full manual annotation.
  - PropertyQualifier2:
    - C25251 : Primary: Occurring first in time or sequence; original; of greatest rank or importance or value.

Using PIR Grid Service

- Retrieve the proteins for gene “BRCA2” (Breast Cancer Gene 2)

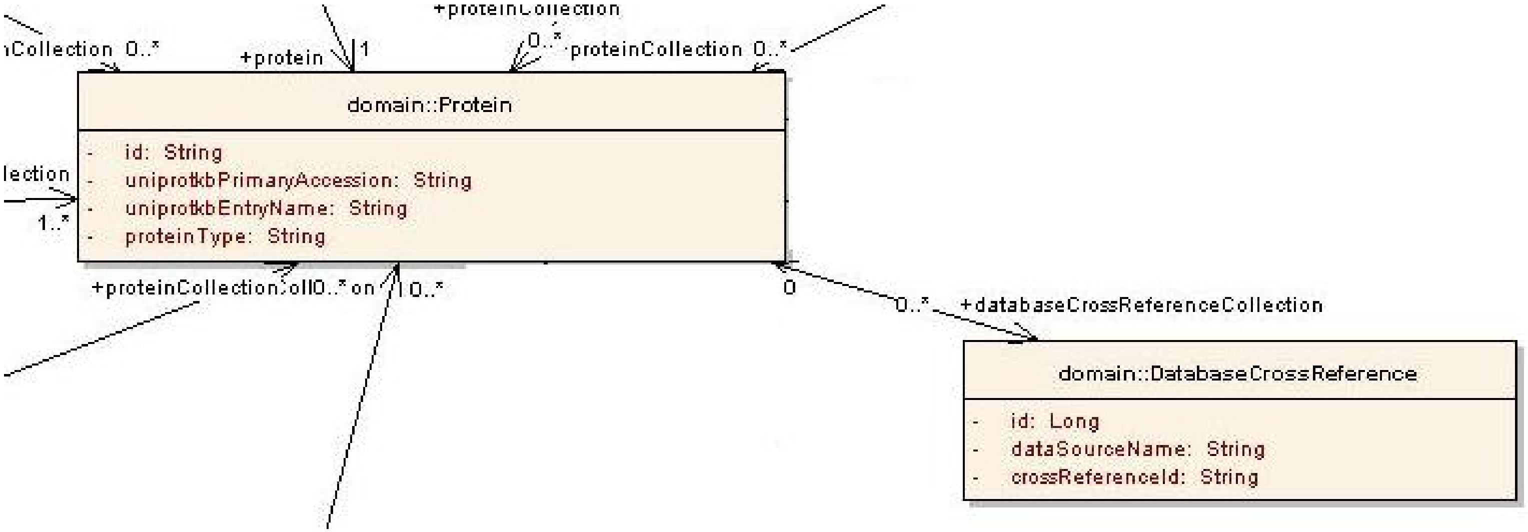
```
<caBIGXMLQuery name="testGene2Protein">
 <Target name="edu.georgetown.pir.domain.Protein“>
   <Objects name="edu.georgetown.pir.domain.Gene">
      <Property name=“name“ predicate="equal” value=“BRCA2"/>
   </Objects>
 </Target>
</caBIGXMLQuery>
```



Using PIR Grid Service

- Find all the proteins that contain the domain “BRCA2 repeat” (PFAM:PF00634, a domain in Breast cancer type 2 susceptibility protein)

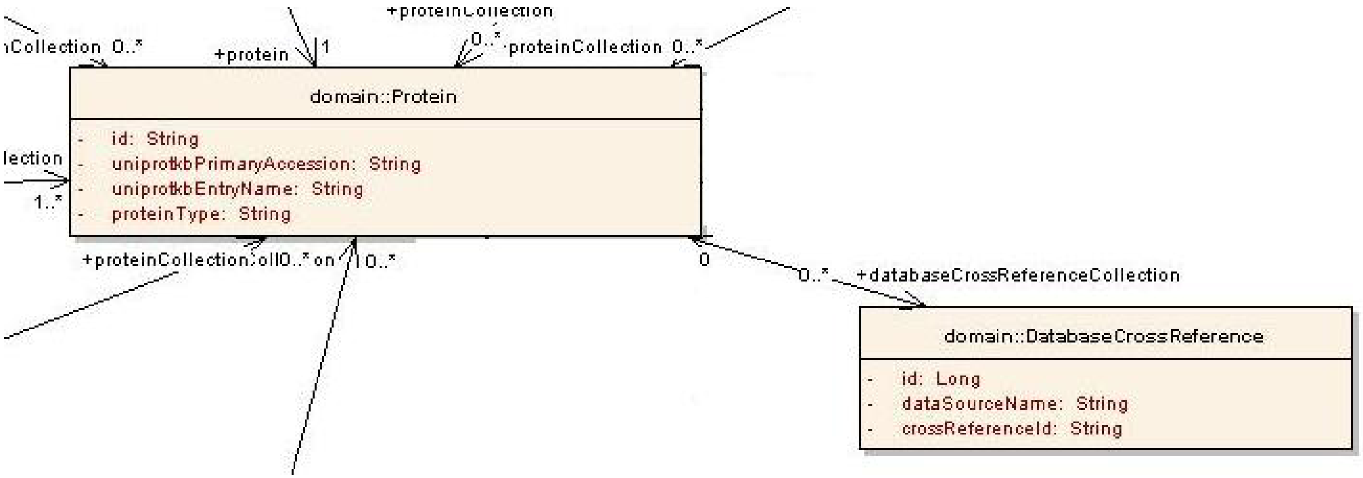
```
<caBIGXMLQuery name="testPfam2Protein">
  <Target name=“edu.georgetown.pir.domain.Protein”>
       <Objects name="edu.georgetown.pir.domain.DatabaseCrossReference">
          <Property name="crossReferenceId" predicate="equal" value="PF00634"/>
       </Objects>
    </Target>
</caBIGXMLQuery>
```


+proteinCollection
nCollection 0..*
+protein
1
0..*
proteinCollection 0..*
domain::Protein
- id: String
- uniprotkbPrimaryAccession: String
- uniprotkbEntryName: String
- proteinType: String
lection
1..*
+proteinCollection:oll0..* on
10..*
0
0..*
+databaseCrossReferenceCollection
domain::DatabaseCrossReference
- id: Long
- dataSourceName: String
- crossReferenceId: String

Using PIR Grid Service

- ID mapping: Find all the database cross-references from various databases corresponding to RefSeq Accession NP_061820

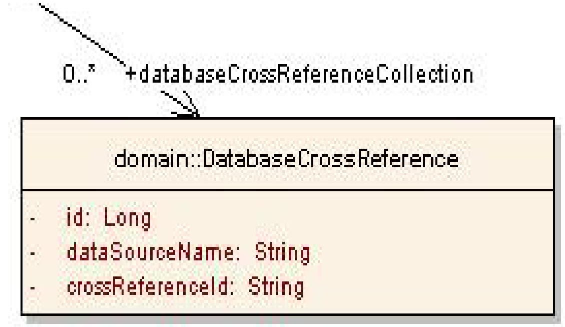
```
<caBIGXMLQuery name="testIDMapping">
 <Target name="edu.georgetown.pir.domain.DatabaseCrossReference“
     path=“edu.georgetown.pir.domain.Protein”>
   <Objects name="edu.georgetown.pir.domain.DatabaseCrossReference">
    <Property name="dataSourceName" predicate="equal"  value="RefSeq"/>
    <Property name="crossReferenceId" predicate="equal" value="NP_061820"/>
   </Objects>
</Target>
</caBIGXMLQuery>
```


+proteinCollection
ıCollection 0..*
+protein
1
0..*
proteinCollection 0..*
domain::Protein
- id: String
- uniprotkbPrimaryAccession: String
- uniprotkbEntryName: String
- proteinType: String
lection
1..*
+proteinCollection:ollO..*on
10..*
0
0..*
+databaseCrossReferenceCollection
domain::DatabaseCrossReference
- id: Long
- dataSourceName: String
- crossReferenceId: String

Discussion

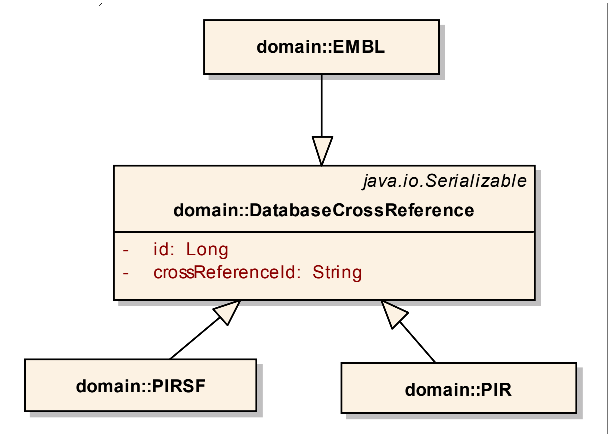
- caBIO and PIR databaseCrossReference objects and dataSourceName – Flexibility vs. better semantics

Current



0..*
+databaseCrossReferenceCollection
domain::DatabaseCrossReference
- id: Long
- dataSourceName: String
- crossReferenceId: String

Future (?)



domain::EMBL
java.io.Serializable
domain::DatabaseCrossReference
- id: Long
- crossReferenceId: String
domain::PIRSF
domain::PIR